**PROJECT SUMMARY** “**ABI Innovation: Cross Species Network Inference**”

**1. Senior personnel**

**PI:** Dennis Shasha (NYU Courant Institute of Mathematical Sciences)

**Co-PIs:** Gloria Coruzzi & Manpreet Katari (NYU Biology, Center for Genomics & Systems Biology)

**Senior Personnel:** Arthur Goldberg (NYU Courant Institute of Mathematical Sciences)

**Collaborators:** Rodrigo Gutierrez, Catolica Universita de Chile.

**2. Intellectual merit of the proposed activity** We propose to develop a Cross Species Network Inference (CSNI) platform that will enable plant biologists to enable researchers to predict how an interacting network of genes/products in new genomes will react *as a system* in response to genetic modifications, ultimately for agricultural, bio-fuel, and ecological benefit. To implement this ambitious goal, tools will integrate genome-scale data acquired in a Target Genome into *inferred* gene networks with the aid of experimentally validated data (e.g. metabolic, protein interaction, etc.) from Reference Genomes. The result will generate a set of testable hypotheses about gene networks in a target genome, as well as suggestions for future experiments, especially time series experiment. This project will leverage the facilities of the current VirtualPlant software platform ([www.virtualplant.org](http://www.virtualplant.org)) developed under an NSF Arabidopsis 2010 Grant (DBI-0445666) including Arabidopsis multinetwork data, analysis, and manipulation tools {Katari, 2009 #1}. As output, we will provide a pipeline of tools for Cross Species Network Inference to the community via a website ([www.CrossSpecies.org](http://www.CrossSpecies.org)) and the NSF*iPlant*Project (see letter). In addition, our CSNI framework will build on the infrastructure of a generic bioinformatic analysis platform engine such as Taverna {Oinn, 2004 #2}, Kepler {Altintas, 2004 #3}, or Galaxy {Blankenberg, #4}. As a proof-of-principle, we will apply this Cross-Species Network Inference framework to predict metabolic and kinase interaction edges. We will then extend CSNI to other genomes, for which experimental data supporting network edges is not yet available. This work will achieve one of the main goals of Systems Biology – predicting network states under untested conditions.

**We divide the work into three aims:**

**Aim 1.** **Inference and validation of an interaction network in Rice as a proof-of-principle.**

Using a known validated network in Arabidopsis (e.g. metabolic, protein:protein, and other validated interactions), homology between Arabidopsis and Rice genes, and transcriptome correlation data in Rice, we will create *inferred* interaction networks in Rice. To validate and refine the approach, these *inferred* Rice networks will be compared with experimentally validated networks from Rice. This can extend to other species.

**Aim 2. Inference of Regulatory Networks Using Time Series** As time series constitute a particularly informative method for inferring networks, we show the power of a model resulting from a closely spaced time series of an N-treatment experiment in Arabidopsis *even in the absence of other data*. We describe our “State-Space Analysis” machine learning method and validate the resulting network based on its predictive accuracy on out-of-sample data.

**Aim 3. Develop A Bioinformatic Pipeline for Cross-Species Inference (CSNI).** Provide a biologist-friendly CSNI software platform (www.CrossSpecies.org) that will infer networks in a target species, given experimentally validated networks in a reference species, homology information and experimental data in the target species.

**3. Broader impacts of the proposed research** This project is the result of a long-standing and highly successful collaboration between biologists at NYU and elsewhere, and computer scientists at NYU's Courant Institute of Mathematical Sciences. The systems biology tools and pipelines resulting from this project will empower biologists to use genomic data to predict a spectrum of gene networks in biology with broad applications to agriculture and the environment. In addition to scientific results, this collaboration extends to joint training of post-docs and graduate students in Systems Biology.

**PROJECT DESCRIPTION**

**RELEVANCE AND JUSTIFICATION TO THE STATED GOALS OF THE ABI/INNOVATIVE**

Our proposal is novel in its framework for cross-species network inference and in its algorithmic ideas for network inference from time series. Its draws its impact from the fact that many plant species will soon be sequenced and so it is of interest (i) to apply information gathered from other species and (ii) to design informative experiments and analyze them through time series. The approach and pipeline tools we develop will be deployed on a biologist-friendly Web interface that plant scientists can use to infer regulatory networks for any crop or trait of interest. Our project addresses several of the ABI goals:

1. *New algorithms for network inference:* Cross-species network inference methods (Aim 1) and Inference from time series data using state space modeling (Aim 2)

2. *Heterogeneous data*: Use of homology and expression to infer metabolic, kinase, and protein-protein networks (Aim 1).

3. *Tools for biological work-flows:* Cross Species Network Inference (CSNI) Pipeline (aim 3).

4. *Enhance education, training and outreach*: Training in Plant Systems Biology

5. *Broaden societal impacts of Systems Biology*: Enable *in silico* predictions for modifying traits of

agronomic and/or environmental value.

**RESULTS FROM PRIOR NSF SUPPORT** This proposal leverages on the accomplishments of the previous parent NSF grant, “Conceptual Data Integration for the Virtual Plant” (DBI-0445666). The VirtualPlant software platform (www.virtualplant.org) {Katari, 2009 #1} integrates genome-wide data concerning the known and predicted relationships among genes, proteins and molecules, as well as genome-scale experimental measurements. VirtualPlant also provides tools that render multivariate information into visual displays (e.g. networks) to highlight biological implications. We have demonstrated the use of tools embodied in the VirtualPlant system to generate hypotheses that were subsequently experimentally validated {Gifford, 2008 #9;Gutierrez, 2008 #7;Thum, 2008 #10; Wang, 2004 #11; Gutierrez, 2007 #12;Nero, 2009 #13}. The first grant had four goals: integration, visualization, synthesis, and prediction.

**1.** **Integration**: *The Arabidopsis Multinetwork: A systems biology tool for hypothesis generation***.** Our VirtualPlant work included assembling the first multinetwork for Arabidopsis, a first step towards a molecular wiring diagram of the plant cell { Katari, 2009 #1; Gutierrez, 2007 #5}. The Arabidopsis multinetwork in VirtualPlant has 16,562 nodes (of which 13,960 are genes) and 97,423 interactions (Fig. 1B, Table I). The multinetwork enables researchers to interpret transcriptome data in the context of all known sources of interaction including protein, DNA, RNA, etc. In one example, a query against the Arabidopsis multinetwork with 834 N-regulated genes resulted in a sub-network of 369 genes connected by one (or more) “expression correlation edges” {Gutierrez, 2008 #7}. At the top of the resulting list of network TF “hubs” (with 47 connections to targets in the N-regulatory network) was the central clock control gene CCA1, a Myb family transcription factor (TF) {Gutierrez, 2008 #7}. Exploration of the network “neighborhood” surrounding this CCA1 TF hub revealed connections to target genes in N-assimilation (Fig. 1C). Using Arabidopsis lines that over-express 35S::CCA1 and by Chromatin-IP {Gutierrez, 2008 #7}, we showed, using phase response curves, that distinct N-metabolites can advance or delay the circadian phase of CCA1 expression. Thus, we derived and validated the novel hypothesis that N-regulation of CCA1 mRNA expression sets the circadian clock. Other examples of networks derived and validated using the VirtualPlant multinetwork are reported in {Gifford, 2008 #9; Thum, 2008 #10; Nero, 2009 #13}. A complementary tool is GeneMania (Warde-Farley D, Donaldson SL, 2010 Nucleic Acid Research) which generates a hypothesis for gene function based on interactions with other genes and their attributes. For a recent review of various plant multinetwork approaches, see [Moreno-Risueno, 2009 #6).

**2 & 3. Synthesis and Visualization: VirtualPlant’s primary analysis tools and functions:** Below is a list of three exemplary tools deployed through VirtualPlant.

**BioMaps**: BioMaps takes one or more sets of genes and determines which functional terms (GO or MIPS) are statistically over-represented in each set with respect to a background population (e.g. Arabidopsis genome). The output is presented in either a tabular format that can be downloaded to Microsoft Excel or a graphical representation based on the appropriate (e.g. GO) directed acyclic graph.

**Sungear**: Sungear is a visually interactive and biologist-driven exploration of standard questions on many experiments on a genomic scale. Sungear can represent an arbitrary number of experiments/lists, all of their disjoint intersections, and their related ontological terms. The position of a circle and arrows emanating from it indicate the input lists of which it is a subset. The size of a circle is proportional to the number of genes in the intersection of those lists (see {Poultney, 2007 #17}). Many biologists find Sungear to be an extremely powerful and interactive tool for analyzing the interrelationships between sets of genes {Gutierrez, 2007 #57}.

**NetMatch:** NetMatch, a cytoscape plug-in, finds all instances of a query graph (e.g. a network motif) in a larger graph {Ferro, 2007 #18}. New versions compute the statistical significance of the motifs found.

Up and coming tools include **GeneSect** whose purpose it is to take a set of collections of genes and to determine whether any pairwise intersections among those collections are either surprisingly large (against a variety of backgrounds) or surprisingly small. Another new tool under development is a powerful cluster architecture to run some expensive tasks such as correlation and network inference in parallel, which relate directly to Aim 2 of the current application.

**4. Extensions:** We have approached dynamic network modeling by applying a machine learning method called “State Space” analysis to time-series data in Arabidopsis learn regulatory networks {Krouk, 2010 #19;Mirowski #57}. This approach is more fully described in the Research Plan (Aim 2) because it relates this new NSF ABI proposal. Our second goal was to extend VirtualPlant to other species, such as Rice , which we have done (Fig. 2).

**Virtual Plant User Community:**

The VirtualPlant user community currently consists of 635 registered academic and commercial users from 36 countries. Among the 347 registered US users, 181 are from academia and 166 are from companies. Examples of commercial users include: Monsanto, Pioneer, Ceres, Syngenta and Unilever. Other countries that also have many users include: UK (78), Australia (27), Germany (24), Chile (22), France (15), Italy (11), Spain (10), Canada (9), Japan (8), Korea (8). In addition, many anonymous users use VirtualPlant, but cannot store their datasets for later iterative analysis.

**VirtualPlant DB**: The VirtualPlant database contains some of the most commonly used data types including metabolic pathways from KEGG and ARACYC, protein-protein interactions from BIND and Interolog databases, and GeneOntology and Gene annotations from TAIR (see Table I for a complete listing of data sources). The database also contains processed data obtained by analyzing publicly available Microarray experiments obtained from NASC {Craigon, 2004 #14}.

**Software and Data Availability**: VirtualPlant is accessible via the website www.virtualplant.org. Registered users (currently > 630) store their data sets and use many tools to analyze their genomic data such as microarray experiments. The website does not require a password and is available for free when used for non-for-profit purposes.

**Plant Genome Application IOS-1025989: TRMS “Cross species network inference: From models to crops” (January 26, 2010):** This ABI application proposes to build tools to infer networks in newly sequence or under-analyzed species. The tools proposed here constitute the computational portion of a proposal that was previously submitted to NSF Plant Genome which was ranked highly meritorious. *All six reviewers noted that the tools were important, timely and would be of benefit to the entire plant community.* Below are excerpts of reviewer comments related to this point.

**Overall Panel Review**: “The effort to make network inference applicable across plant species is important and timely. There was no doubt the proposed methods would be effective. There is excellent potential for tools from this project to be widely applied. This was seen as a strong proposal from an excellent interdisciplinary team of researchers.”

**Review 1**: “This project proposes to leverage the VP platform to create a pipeline of tools for cross species network inference in plants. This is a highly relevant effort that will benefit many ongoing hypothesis driven projects that lack the tools or capability to include network analysis. The large effort in implementation is well justified as this will be a major resource and wide usability will depend on stability, power and ease of use. I think there will be a lot of “bang for the buck” including novel scientific insights. … Tool development efforts are well integrated in cyberinfrastucture, including iPlant and Galaxy”.

**Review 6**: “Shasha et al propose to develop, validate and deploy an analysis pipeline for comparative inference of gene function and interaction based on similarities in NT sequence, regulatory regions and transcription patterns. Such a tool is sorely needed with the growing number of genome and trancriptome sequences coming available for the emerging model and non-model species. … As such, the proposed development of a web based Cross species network inference database and analysis tool would be a major contribution.”

**Review 3**: “With the emerging genome sequences and functional genomics datasets now available for other plant species, the time has now come to apply the gene network construction and analysis functions within the VP to crop plants.”

**Review 4**: “A resource will be created for the entire scientific community (the cross species network inference pipeline) which will be freely available on the web. This work will…develop a tools that will advance research in many areas of plant biology.”

**Review 5**: “The proposed science is of high quality and internationally competitive. The application area is of the highest importance.”

**Review 2**: “The CSNI tool would likely be used by the wider plant biology community.”

There were some criticisms by the reviewers as well: one pointed out that certain network edges should enjoy more confidence than others. The reviewer suggests that we reflect confidence in weights. Our time series machine learning approach will do that. Another reviewer pointed out that using correlation across all experiments may work less well than choosing experiments carefully depending on the genes of interest. We refine our choice of experiments in Aim 1. Yet another criticism suggested that our techniques for obtaining orthology should be compared with those of InParanoid and OrthoMCL. We will address these criticisms as they come up in the rest of this proposal.

**PUBLICATIONS: Peer reviewed journal articles, chapters and books:**

**VirtualPlant: Tool development for Plant Systems Biology**

Katari M, Nowicki S, Aceituno F, Nero D, Kelfer J, Thompson L, Cabello J, Davidson R, Goldberg A, Shasha D, Coruzzi G, Gutierrez R (2009) “VirtualPlant: A software platform to support Systems Biology research”. **Plant Physiol**. Dec 9 *(Epub ahead of print) (SUZAN UPDATE THIS).*

Nero D, Kelfer J, Katari M, Tranchina D, Coruzzi G (2009) “In silico Evaluation of Predicted Regulatory Interactions in Arabidopsis thaliana”. **BMC Bioinformatics**. Dec 21;10(1):435.

Poultney C, Gutierrez R, Katari M, Gifford M, Paley W, Coruzzi G and Shasha D (2007) “Sungear: Interactive visualization, exploration & functional analysis of genomic datasets”. **Bioinformatics**, 23:259-61.

Ferro A, Giugno R, Pigola G, Pulvirenti A, Skripin D, Bader G, Shasha D, “NetMatch: a Cytoscape Plugin for Searching Biological Networks” **Bioinformatics**, 2007 23(7):910-912.

**Applications of VirtualPlant: Hypothesis Generation and Testing**

Krouk G, Tranchina D, Lejay L, Cruikshank A, Shasha D, Coruzzi G and Gutierrez R (2009) “A systems approach uncovers restrictions for signal interactions regulating genome-wide responses to nutritional cues in Arabidopsis.” **PloS Comp Biol**. Mar;5(3):e1000326. *(Highly Accessed).*

Gutierrez R, Stokes T, Thum K, Xu X, Obertello M, Katari M, Tanurdzic M, Dean A, Nero D, McClung R and Coruzzi G (2008) "Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1" **Proc. Natl Acad Sci USA** 105, 4939-4944. *(Faculty of 1000 recommended: Factor 3)*

Gutierrez R, Gifford M, Poultney C, Wang R, Shasha D, Coruzzi G, Crawford N (2007) "Insights into the genomic nitrate response using genetics and the Sungear Software System" **Journal of Experimental Botany** doi: 10.1093/jxb/erm079

Gutierrez R, Lejay L, Chiaromonte F, Shasha D, Coruzzi G (2007) "Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive biomodules in Arabidopsis" **Genome Biology**, 8: R7. *Faculty 1000 (Must Read: Factor 6)*

**Plant Systems Biology: Reviews, Books and Outreach**

Ruffel S, Krouk G, Coruzzi G (2009). "A Systems View of Responses to Nutritional Cues in Arabidopsis: Towards a Paradigm Shift for Predictive Network Modeling”. **Plant Physiol**. Nov 25 *(epub ahead of print) (SUZAN UPDATE)*

Gutierrez R, Coruzzi G., Eds (2009) Book: “Plant Systems Biology”, **Annual Plant Reviews**; Blackwell Publishing: Oxford, UK, 2009, Vol. 35. 360 pages.

Coruzzi GM, Burga A, Katari MS, and Gutierrez RA (2009) “Systems Biology: Principles and Applications in Plant Research”. In “Plant Systems Biology”, **Annual Plant Reviews**; Blackwell Publishing: Oxford, UK, 2009, Vol. 35. Pgs 3-31. *Book Chapter.*

Gifford M, Gutierrez R, and Coruzzi G (2006) "Modeling the Virtual Plant: A Systems Approach to Nitrogen-Regulatory Gene Networks". Essay 12.2 Chapter 12. Assimilation of mineral nutrients; In **A Companion to Plant Physiology*,*** Fourth Edition, Lincoln Taiz and Eduardo Zeiger, http://4e.plantphys.net/article.php?ch=12&id=352

Gutierrez R, Shasha D and Coruzzi G. (2005) "Systems Biology for the Virtual Plant". **Plant Physiol.** Vol 138, pp 550-554.

**Education & Training**: The development of the Systems Biology tools and the Virtual Plant software platform has trained undergraduates (UG), MS and PhD students in Systems Biology. Students trained include **Undergraduates**: Steve Nowicki (NYU CAS), Varuni Prabhakar (Barnard College), Rebecca Davidson (BS Computer Science); **Masters Students**: Ana F. Arroja (MS student, NYU Courant), Ranjita Iyer (MS Computer Science), Jonathan Kelfer (MS Computer Science), Lee Parnell (MS Computer Science), (Jarod Wang, MS Computer Science); **PhD Students**: Chris Poultney (PhD student, NYU Courant), Aris Tsirigos (PhD student, NYU Courant), Saurabh Kumar (PhD student, NYU Courant). These students have gone on to PhD programs (Prabhakar, Parnell), post-docs (Poultney, Tsirigos) as well as to industry (Kelfer, Bloomberg).

**RESEARCH DESIGN**

**Aim 1: Inference and validation of interaction networks in Rice as a proof-of-principle.**

***Rationale***: This aim will use the data-rich genomic resources of Arabidopsis and Rice to evaluate our Cross Species Network Inference (CSNI) approach by inferring an interaction network in Rice and then validating the network against Rice validated data. The data used in this proof of principle will be: 1) Arabidopsis validated data that includes metabolic pathway data obtained from KEGG {Kanehisa, 2004 #20}, protein interaction data obtained from the Biomolecular Interaction Network Database (BIND) {Bader, 2002 #21} plus other experimentally determined protein interactions {de Folter, 2005 #22;Popescu, 2007 #23}, 2), Rice-Arabidopsis gene homology, and 3) Rice microarray expression data. This *inferred* Rice network will be compared to the *known* validated data for Rice including metabolic data from KEGG and protein:protein interaction data from BIND. Our goal is to use the well-known gene interaction datasets for Arabidopsis and Rice to develop and validate a methodology for inferring networks for other species, and later apply the CSNI pipeline to under-analyzed species that have little interaction information (e.g. Medicago). See Fig. 3 for overall plan. Our approach builds on inference approaches based on expression (Gholami, Fellenberg, 2010 Bioinfroamtics, “Cross-species common regulatory network inference without requirement for prior gene affiliation”), homology (Haiyuan Yu, Nicholas M Luscombe, Hao Xin Lu, Xiaowei Zhu, Jing-Dong J. Han, Nicolas Bertin, Sambath Chung, Marc Vidal, Mark Gerstein (2004) Genome Res), expression and homology (Mutwil , Persson, 2008, Nucleic Acid Research, “GeneCat – novel webtools that combine BLAST and co-expression analyses”), and also based on integration of several different types of associations ( Geisler-Lee, O’Toole et al 2007, Bioinformatics) (Mutwil et al 2008)

**Step 1.** **Obtain a reference validated Arabidopsis interaction network based on experimentally supported data**. For our validated Arabidopsis networks, we have assembled metabolic interactions (KEGG; 19,688 interactions) {Kanehisa, 2004 #20}, protein:protein interaction data from BIND (949 interactions) {Bader, 2002 #21}, protein-chip interaction data for MADS box (272 interactions) {de Folter, 2005 #22} and protein chip interactions for Calmodulin (755 interactions) {Popescu, 2007 #23}. Although we refer to the metabolic and protein interactions data as validated, many of the pathways in the KEGG and AraCyc databases are based on computational predictions, while 25% are validated in the literature {Masoudi-Nejad, 2007 #28;Zhang, 2005 #29}.

**Step 2. Identify Rice homologs of Arabidopsis interaction pairs.** Connect every gene in the Arabidopsis interaction network with its Rice homologs. This technique can employ various homology methods, distance or parsimony based. In our preliminary analysis (Table II), we analyzed one-to-one homology by obtaining reciprocal top BLAST pairs. We also used distance-based BLAST with an e-value cutoff of E-20 to capture one-to-many homology relationships {Tatusov, 1997 #33} which captures the gene duplication events prevalent in plant genomes {Zhang, 2003 #34}.

**Step 3. Build a Rice correlation network based on publicly available Rice microarray expression experiments.** We downloaded 32 Rice microarray experiments from GEO {Barrett, 2006 #35}, log transformed the MAS5 {Pepper, 2007 #36} normalized values, and Pearson correlated all pairs of the genes whose measurement variances (after normalization) lie in the upper 20% of all genes. We inferred a correlation edge between gene pairs whose expression vectors were significantly correlated (p-value <0.05, meaning less than a 5% chance of a non-zero correlation by chance) and correlation value > 0.5 or >0.7 (Table II).

**Step 4.** **Build an *inferred* Rice network**. A pair of Rice genes in our *inferred* Rice network may be connected by no edge, by only an expression correlation edge, by only a homology edge from Arabidopsis, or by both of these types of edges. We focus on Rice gene pairs connected by both types of edges – homology and correlation. This network is called the *inferred* Rice network.

**Step 5. Obtain a reference validated Rice network that contains edges representing known interactions.** Our initial Rice validated network was constructed from 10,976 metabolic interactions and 334 protein-protein interactions for Rice from KEGG {Kanehisa, 2004 #20} and BIND {Bader, 2002 #21}, respectively.

**Step 6. Evaluate *Inferred* Rice Network:** This step computes the similarity and p-value (significance) between the *inferred* and validated Rice networksby using a network intersection tool called ***NetSect*** which is described below. We evaluated the quality of each subset of edge types in the inferred network.

***NetSect*: Evaluating the Accuracy of the *Inferred* Network**. Given networks *N* (“inferred”) and *M* (validated), with edges *E(N)* and *E(M)* respectively, one can measure their similarity by computing *size( intersection( E(N), E(M) )) / size(union( E(N), E(M) ) )*, which equals *1* when *E(N)* and *E(M)* are identical and zero when they are disjoint. We will also compute the recall and precision of the *inferred* network’s ability to predict edges in the reference validated network. To compute a p-value for the *inferred* network's reconstruction of the reference network, ***NetSect*** computes the similarity of the inferred and validated networks and then computes a p-value by comparing the sample similarity with the similarity of a collection of random networks having the same topology (i.e. isomorphic) as the inferred network, with vertices drawn from the entire genome. This use of randomness corresponds to the null hypothesis that the inferred network is no better than a random choice of edges.

We suggest two main conclusions from our preliminary analysis of Cross Species Network Inference (Steps 1-6 above) shown in Table II. *First*, homology alone does an excellent job of creating an inferred network for metabolic edges. Of the 3,594 edges in the Rice metabolic network inferred via reciprocal top hits, 52.4% or 1,883 are validated in the Rice validated KEGG metabolic interactions. *Second*, correlation significantly enhances the prediction of the inferred Rice network. Of the 387 inferred Rice metabolic interactions predicted with the intersection of homology (reciprocal top hit) and correlation (>0.7), 275 inferred interactions (or 71.1%) are validated by the Rice metabolic validated network, which is a statistically significant improvement in precision (p-value < 0.001) (Table II). Based on ***NetSect*** analysis, the predictive power of the reciprocal top-hit inferred Rice metabolic network is significant (p-value < 0.001), with or without expression correlation data. The precision of this prediction is so high, that we hypothesize that many of the remaining 29% of predicted edges may represent true interactions that are currently missing from the Rice KEGG metabolic database.

*Kinase conclusions go here*

**Step 7. Expand validated and network inference into a “multinetwork” containing multiple edge types**. We will use techniques analogous to Steps 1-6 to infer networks based on other edge types. For example, we will add regulatory interactions including protein🡪DNA (AGRIS: 343 interactions) {Davuluri, 2003 #24} and miRNA:RNA interactions {Griffiths-Jones, 2006 #25;Gustafson, 2005 #26;Lu, 2005 #27}. Expanding the validated networks to include these datasets will enable us to create an inferred multinetwork that includes: protein:protein, Protein:DNA, miRNA-RNA and Metabolic edges.

**Parameter optimization.** As one would expect, the choice of data sources and homology algorithms and parameters for CSNI, greatly influences the accuracy of the inferred Rice networks. To simplify the selection of these parameters for biologists, we will systematically explore the space of these inputs, with the objective of maximizing the accuracy of our network inference predictions. A well-known technique for finding globally optimal parameters is “*simulated annealing*”, which is a probabilistic heuristic for finding global minima in large search spaces {Michaelewicz, 2004 #37}. Ideally, the experiments used for gene expression correlation will include many different developmental stages, different organs, and different biotic and abiotic treatments such as the ones just recently released for Rice on GEO NCBI {Wang, 2009 #38}.

**Expected Outcomes of Aim 1**: We will expand this cross species network inference and validation analysis to include other homology methods (e.g. parsimony-based homology {Chiu, 2006 #39} and other methods like COGS {Tatusov, 2000 #40}, InParanoid {Berglund, 2008 #41}, for example). We will also expand our data sources to larger datasets for expression and protein interaction from a variety of species as they become available. For example, currently there are large scale Arabidopsis and Rice protein interaction datasets being created (Joe Ecker – personal communication) which will allow us to better test and refine our methods (and the methods of ( Geisler-Lee, O’Toole et al 2007, Bioinformatics)) to be used on newly sequenced species This Aim provides a testing ground and validation for the CSNI pipeline approach that we will automate in Aim 3.

**Aim 2: Inference of regulatory networks: Develop a time series expression methodology using multiple data sources to infer regulatory networks.**

***Rationale:*** A principal aim of the biological enterprise is to learn which genes affect which functionalities and how. Whereas it is possible to create models and draw networks based on correlation and other data-driven relationships, the ultimate test of any model is the ability to predict the values of assays (e.g. expression levels) on conditions that were not used in training. If the model is in the form of a causal network, it can be used to suggest gene modifications that may optimize the performance of a species to some useful end. Because causality moves forward in time, time series experiments are a particularly promising source of network structure. In this aim, we describe our methodology in some detail, explain prior results, and explain how we integrate the methodology with other existing information (e.g. validated transcription factor binding sites, structurally based contact-binding sites, and expression data). The idea is to use this methodology for time series experiments in newly sequenced/under-analyzed species. The methodology runs on a standard parallel cluster, so our cluster management software (not described) can be used for any site that downloads Virtual Plant.

**Predicting Arabidopsis regulatory networks using time series data and “State Space” analysis a machine learning approach.** The experimental approach of our Arabidopsis time-series {Krouk, 2010 #19} was to monitor transcriptome responses to NO3- treatment at 0, 3, 6, 9, 12, 15 and 20 min, using ATH1 chips. In order to build a regulatory network that could predict these TF-target interactions, we used a machine learning method, “State-Space” modeling to generate predictions for regulatory networks {Mirowski, 2009 #56}. The State-Space model synthesizes Bayesian and Markovian approaches (in which each gene’s expression value at a time *t* is assumed to depend directly only on the state of potentially all the genes at the previous time point and indirectly on values from previous time points). {Murphy, 1999 #42; Mirowski, 2009 #56}.

In the “State Space” model depicted in Fig. 5B, each node represents the values of all gene expressions at a particular time point. Typical values of all gene expressions are depicted as a heat map in Fig. 4. The goal of this approach, is to *learn* the function that determines the change in expression of a target gene z*j*, as a linear (or if needed non-linear) combination of the expression of a relatively small number of transcription factors (typically up to three or four) (Fig. 4). As applied to our problem, the set of all genes at time t is modeled by a “latent” (i.e. hidden) variable (denoted Z(*t*)) from which noisy and sometimes missing observations Y(*t*) are made. Latent variables are represented by large red circles, and observed variables by large black circles in Fig. 5B. The relationship between latent and observed variables is the identity function *h* with added Gaussian noise (represented by a black square in Fig. 5B). An unknown *function f* (represented by a red square in Fig. 5B) relates the values of latent variables Z(t) and Z(t+1) (for all *t)* corresponding to consecutive time measurements as a Markov chain. The *dynamical function f* factors in both transcription factors and their target genes (e.g. other TFs or target genes), as shown in Fig. 4. The “State Space” learning algorithm iteratively infers the function of latent values of transcription factors that determines the changes to target genes. Learning the *function f* corresponds to finding parameters of *f* that minimize the prediction error while penalizing functions that are excessively complex (i.e. require many transcription factors to determine the change in expression of a target).

An iterative procedure tries to learn the dynamical relationship between latent gene expression variables **z**(*t*) while maintaining the latent variables **z**(*t*) as close as possible to the observed Affymetrix measures **y**(*t*). The algorithm consists in a) minimizing the sum of quadratic errors of the dynamical and the observation models with respect to the latent variables **Z** by using gradient descent on the latent variables {Mirowski, 2009 #56} (this is the inference step); and in b) minimizing the sum of quadratic errors of the dynamical model using conjugate gradient, LARS {Efron, 2004 #58} or Elastic Nets {Zhu, 2005 #59}optimization on the parameters of **F** (this is the learning step). During the learning step, sparse gene regulation networks are obtained by penalizing dense solutions using L1-norm regularization, which amounts to adding a *λ*-weighted penalty to the dynamical error term, as in the LASSO initially described by {Tibshirani, 2006 #60}.

To test the ability of the “State Space” approach to generate a *predictive* regulatory network, we built a regulatory network using the Arabidopsis time-series data up to 15 minutes (training set: 0, 3, 6, 9, 12, 15 min) and used the resulting network to *predict* the direction of gene change (up regulation or down regulation) from 15 min to 20 min (Fig. 5). Our State Space predictions of gene regulation were correct for 74% of the genes in a small network of 76 genes(Fig. 5B). As a basis for comparison, the *"naive trend forecast"* that predicted the direction of change from 15 to 20 min to be in the same direction as the movement from 12 to 15 min, was correct for only 52% of the genes, just slightly better than random (Fig. 5C), p-value < 0.006. This “State Space” model can also be used to predict the “most influential TFs” in the network (e.g. the one that is predicted to influences the most genes in the network), and to generate a time-dependent regulatory network model for the control of N-assimilatory pathway genes.

10

When compared with other approaches {Bonneau, 2007 #61;Bonneau, 2006 #50}, {Shimamura, 2009 #62}, and {Wang, 2006 #63}, our method showed a slight improvement in accuracy and had a better signal to noise ratio. Further the method reduces the importance of initial parameters by using random starting points and bootstrapping, thus offering a principled way to deal with uncertainty and avoid over-fitting in micro-array measurements. Further, our method easily allows the addition of “hints” in the form of known transcription factor-binding relationships. Finally, the method generalizes to larger networks: in a network of 550 genes (67 TFs, 483 non-TF genes) including the original 76, the state-space model (running on top of the clustering algorithm CMonkey {Reiss DJ, Baliga NS, Bonneau R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. \*BMC Bioinformatics\* (2006)} using default settings) predicted the direction accurately for 67.7% of all the genes in the 15-20 minute time point compared with 51% for the naïve trend forecast prediction.

Here is a summary of the approach

1. Find sentinel genes for the given treatment.
2. Test them at a variety of times
3. Decide on the times of global expression experiments
4. Do enough experiments
5. Analyze using our machine learning algorithm
6. Try a leave-out-one test. If resulting network gives good enough predictions, then use it.

**Expected outcomes of Aim 2:** The results of this aim will generate a robust package that can be used across species and for a variety of time series experiments. The goal will be to help the design of time series and to construct regulatory networks using a high performance parallel cluster architecture. From this, we can identify a core set of regulatory networks.

**Aim 4: A Bioinformatic Pipeline for Cross-Species Network Inference (CSNI).**

***Rationale*:** In this aim, we will build a publicly available, production quality, Cross-Species Network Inference (CSNI) pipeline that will provide the plant scientist community (especially those with no informatics training) with a biologist-friendly tool for inferring gene networks in newly sequence/unanalyzed species. CSNI employs data about two species, 1) the under-analyzed species – which we call the *target* species, and 2) a species that has been deeply studied, which we call the *reference* species. The basic idea of CSNI is that the larger data set from the reference species will be mapped by homology into the target species, and combined with data about the target species to infer a network for the target species.

Figure 3A illustrates the CSNI pipeline. A plant scientist who wants to infer a gene network for a target species will set the free parameters that determine the homology and inference methods of CSNI. These include i) the pair of reference and target species chosen, ii) the data sets selected from these species, iii) the homology mechanism and its parameters (such as BLAST E-value thresholds, COG {Tatusov, 2000 #40}, InParanoid {Berglund, 2008 #41, if distance-based homology is desired or parsimony methods {Chiu, 2006 #39}), and iv) the *Inference* rules which combine these data into the target species' inferred network.

In the inference step, the biologist chooses a validated network (e.g. metabolic, protein-protein, etc.) in the reference species. Next, the biologist chooses some parameter settings or allows an optimization technique such as simulated annealing to set those parameters. For example, a combination rule might infer a regulatory edge in the target species if the edge’s genes were connected by an expression edge with correlation > 0.7 and the edge had homologous genes connected by a regulatory edge in the reference species’ validated network. Given all this, CSNI infers the target network.

plan to deploy CSNI ([www.CrossSpecies.org](http://www.CcrossSspecies.org)) on several platforms, first on our VirtualPlant website (www.virtualplant.org), and second on *iPlant* (see S. Goff letter). We will also deploy CSNI on one of the widely-used bioinformatic workflow engines: Taverna {Oinn, 2004 #2}, Kepler {Altintas, 2004 #3} or Galaxy {Blankenberg, #4}. Implementing the CSNI pipeline on top of one (or more) of these bioinformatic workflow engines is important because they provide increasingly popular platforms for developing computational genetic analyses, and provide generic support for reproducible bioinformatic analyses.

**Expected outcomes of Aim 3.** The CSNI pipeline analysis constructed in Aim 3, and made available to the community as a biologist-friendly interface, will empower plant biologists to use network approaches to derive testable hypothesis for gene functions in crop species for which limited genomic information is available. Identifying networks conserved between reference and crop species will also enable researchers to focus their translational studies from models to crops.

**TIMELINE:**

**Year 1:** Aim 1. Extend cross species network inference using validated protein:protein and metabolic interaction networks for Rice and Arabidopsis to other homology methods. Extend network inference analysis beyond protein-protein interaction to validated regulatory (AGRIS) edges as well as miRNA-RNA edges. Aim 2: Analyze time series experiments in various plant species to validate our network inference approach. Aim 3: Assemble validated networks in the 3-5 target crop species beginning with Medicago, Corn, Grape. Select bioinformatic workflow platform on which we will deploy.

**Years 2-3:** Aim 2. Make the State Space analysis platform available to the community, including facilities to suggest needed experiments in under-analyzed species. Aim 3. Deploy the first version of the CSNI analysis pipeline for cross species network inference to collaborators (D. Cook, U Davis; R. Gutierrez, Chile), including facilities to do parameter optimization using heuristic techniques like simulated annealing and genetic algorithms.

**Years 4-5:** Apply the computational pipeline to infer networks in several crop species for example corn and grape. Deploy the full computational CSNI pipeline for cross-species network inference to plant community via CSNI (www.CrossSpecies.org) linked to VirtualPlant, iPlant and a selected workflow platform (e.g. Galaxy). Make all software available as a webservice.

**PLAN TO INTEGRATE RESEARCH AND EDUCATION**.

**Cross training of Biologists and Computer Scientist in Systems Biology**. The development of Systems Biology tools in this project has and will involve biologists teaching computer scientists about topics like genetics, experimental genomics, and the computational challenges of analyzing genomic data. We do this informally at our weekly joint lab meetings at which graduate students and post docs from NYU Biology and NYU Courant each present their work to the group. This project involves a team of three resident full time computer scientists working within a biology lab, interacting closely with wet bench biologists. The senior computer scientists (Shasha, Katari and Goldberg) are also involved in training and engaging computer scientist students at all levels in the emerging field of Systems Biology. In the last six months, they have trained two PhD students, two interns and two MS students from Courant working in this environment. For a complete listing of students trained in the past 4.5 years, see Education and Training section in Results from Prior support.

**Workshops and Classroom Training in Genomics and Systems Biology**: We also provide formal training in the form of workshops and classes to enable Systems Biology. Examples of this include a weekly software workshop in “R”, which aims to teach biologists how to analyze their own genomic data. A workshop on VirtualPLanthas been taught two times, once by Jonathan Kelfer, a MS student working on the project and most recently by Manrpeet Katari, co-PI. Students have included several faculty on sabbatical at NYU including most recently: MaryLou Guerinot and Rob McClung of Dartmouth. Students will be exposed to Genomics and Systems Biology also through a series of formal courses offered by faculty at NYU’s Center for Genomics and Systems Biology including: G23.1128 Systems Biology; G23.1130 Applied Genomics: Introduction to Bioinformatics & Network Modeling; G23.1127 Bioinformatics & Genomes. PhD students have and will continue to present their work in the weekly PhD seminar series hosted by the Biology Department. Computational students will be involved in constructing the pipeline and making it perform through the use of parallelization. Such students will also help to develop and test optimization and machine learning algorithms for network inference.

**Training Postdocs as educators**. In this project, Post-Docs are paired up with graduate students, undergraduate students, and technicians in the laboratory to practice mentoring skills in a research context. At NYU, post-docs are also afforded the opportunity to teach and are mentored by faculty advisors. Post-Docs also receive counseling from their co-mentors and practice presentation skills during regular group-lab meetings, through a Post-Doc seminar series, and at annual poster sessions at NYU.

**PLAN TO INTEGRATE DIVERSITY**

W we are committed to training scientists at the graduate and postdoctoral levels who can do independent research that cuts across fields and expertise in evolutionary genomics.  Our research team is also committed to diversity.  Researchers in our previous Plant Genome grant included Hispanic and African-American students.  We will continue to actively seek out and recruit scientists from under-represented minorities to participate in our research in our continuing commitment to increase diversity in our research program. Five female scientists are associated with this project: Coruzzi (co-PI); Rebecca Davidson (Programmer); Varuni Prabhakar (UG Programmer); Ana Arroja (MS); Ranjita Iyer (MS Courant). Damion Nero a minority PhD student has written programs contributed to the Virtual Plant project.

**SHARING OF RESULTS**

The informatic analysis pipelines for Cross Species Network Inference (CSNI), discussed in Aim 3 will be made available to the community free of charge, deployed on a website (www.crossspecies.org) linked to several additional platforms, first to VirtualPlant website (www.virtualplant.org), and second to *iPlant* (see S. Goff letter), and third as a webservice. **Publications:** The results of our analysis of the data we generate will be made available through peer- reviewed literature as it is the most appropriate way to make this information available.

**MANAGEMENT PLAN**

To coordinate and facilitate interactions between individuals, Dennis Shasha (NYU Computer Science) will serve as the overall Project Manager and Gloria Coruzzi (NYU Biology) will serve as a biological advisor and conduit to a working lab and the wider plant community. The role of the Project Manager is to oversee the daily operations of the project and ensure that the needs and concerns of the participants are addressed on a day-to-day basis between the participants involved. The project manager will also facilitate communication between PIs, post-docs, graduate students and laboratory technicians by scheduling weekly meetings of all participants to manage immediate issues regarding research needs. We will also schedule day-long meetings twice a semester with collaborators on the computational end (Rodrigo Gutierrez, Chile), to do evaluation of work status and long term planning.

**Bioinformatics manager:** Dr. **Manpreet Katari** (NYU Biology) will be in charge of the bioinformatics data. To enable efficient information exchange of raw and processed data, a file server has been set up at the NYU to store and distribute data and its analysis among users at NYU Biology and NYU Courant. Dr. Katari will maintain the web server, database server, and the multinetwork database.

**Software development manager: Dr. Arthur Goldberg** (NYU Courant) will manage the development of new software analysis tools and pipelines to enable Cross Species Network Inference (CSNI) which will support the different species and inference, and also new pipelines for cross species analysis, especially as they relate to crop species in coordination with the PI, the programmer Rebecca Davidson, and a computer science doctoral student.

**Website:** We have set up a web site to house the development of Cross Species Network Inference tools and pipelines, which is accessible at: www.CrossSpecies.org

**Principal Investigators:** Shasha and Coruzzi will each commit a minimum of a half-day per week to this project. This will include supervision of personnel, organizational meeting attendance, and intellectual developments and contributions.

**Role of senior participants and timeline:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Institution** | **Role** | **Aim** |
| ***Dennis Shasha***  PI | NYU Courant | Project Leader | Oversee Aims 1, 2, 3 |
| ***Gloria Coruzzi,***  Co-PI | NYU Biology | Co-leader: Biological | Oversee Aims 2 & 3 |
| ***Manpreet Katari***  Co-PI | NYU Biology | Bioinformatics Manager | Aims 3 |
| ***Arthur Goldberg***  Senior Personnel | NYU Courant | Software developer | Aims 1 |
| ***Rodrigo Gutierrez***  Consultant | U Catolica,  Chile | Assembling validated networks for targets | Aim 1 |

**COORDINATION WITH OUTSIDE GROUPS**

**Please see attached letters of collaboration:**

**Rodrigo Gutierrez (U Catolica, Chile)** Dr. Gutierrez, the creator of the Arabidopsis multinetwork (Gutierrez et al 2007) will assist in the assembly of multinetworks for crop species including Vitis, Corn and Medicago.

**iPlant (see letter from iPlant Project Director, Steve Goff)** We will coordinate with iPlant to make our Cross species network inference platform (CSNI) modular, independent and accessible with and compatible with iPlant, and accessible using other annotation analysis platforms such as Galaxy and Taverna. We will also make our currently developed VirtualPlant tools accessible to iPlant, as per letter by (S. Goff).

**REFERENCES CITED**

1. Katari, M.S., et al., *VirtualPlant: A software platform to support Systems Biology research.* Plant Physiol, 2009.

2. Oinn, T., et al., *Taverna: a tool for the composition and enactment of bioinformatics workflows.* Bioinformatics, 2004. **20**(17): p. 3045-54.

3. Altintas, I., et al., *Kepler: an extensible system for design and execution of scientific workflows.* Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004: p. 423--424.

4. Blankenberg, D., et al., *Galaxy: a web-based genome analysis tool for experimentalists.* Curr Protoc Mol Biol. **Chapter 19**: p. Unit 19 10 1-21.

5. Gutierrez, R.A., et al., *Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis.* Genome Biol, 2007. **8**: p. R7.

6. Moreno-Risueno, M.A., W. Busch, and P.N. Benfey, *Omics meet networks-using systems approaches to infer regulatory networks in plants.* Curr Opin Plant Biol, 2009.

7. Gutierrez, R.A., et al., *Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1.* Proc Natl Acad Sci U S A, 2008. **105**(12): p. 4939-44.

8. Nero, D., et al., *In silico evaluation of predicted regulatory interactions in Arabidopsis thaliana.* BMC Bioinformatics, 2009. **10**: p. 435.

9. Gifford, M.L., et al., *Cell-specific nitrogen responses mediate developmental plasticity.* Proc Natl Acad Sci U S A, 2008. **105**(2): p. 803-8.

10. Thum, K.E., et al., *An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in Arabidopsis.* BMC Syst Biol, 2008. **2**: p. 31.

11. Wang, R., et al., *Genomic analysis of the nitrate response using a nitrate reductase-null mutant of Arabidopsis.* Plant Physiol, 2004. **136**(1): p. 2512-22.

12. Gutierrez, R.A., et al., *Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis.* Genome Biol, 2007. **8**(1): p. R7.

13. Nero, D., et al., *A system biology approach highlights a hormonal enhancer effect on regulation of genes in a nitrate responsive "biomodule".* BMC Syst Biol, 2009. **3**: p. 59.

14. Craigon, D.J., et al., *NASCArrays: a repository for microarray data generated by NASC's transcriptomics service.* Nucleic Acids Res, 2004. **32**(Database issue): p. D575-7.

15. Ideker, T., et al., *Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network.* Science, 2001. **292**(5518): p. 929-934.

16. Gutierrez, R.A., D.E. Shasha, and G.M. Coruzzi, *Systems Biology for the Virtual Plant.* Plant Physiol., 2005. **138**(2): p. 550-554.

17. Poultney, C.S., et al., *Sungear: interactive visualization and functional analysis of genomic datasets.* Bioinformatics, 2007. **23**(2): p. 259-61.

18. Ferro, A., et al., *NetMatch: a Cytoscape plugin for searching biological networks.* Bioinformatics, 2007. **23**(7): p. 910-2.

19. Krouk, G., et al., *High resolution dynamic transcriptome of Arabidopsis roots in response to nitrate: Molecular physiology and predictive modeling.* submitted, 2010.

20. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic Acids Res, 2004. **32 Database issue**: p. D277-80.

21. Bader, G., D. Betel, and C. Hogue, *BIND: the Biomolecular Interaction Network Database.* Nucleic Acids Res., 2002. **31**: p. 248.

22. de Folter, S., et al., *Comprehensive interaction map of the Arabidopsis MADS Box transcription factors.* Plant Cell, 2005. **17**(5): p. 1424-33.

23. Popescu, S.C., et al., *Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays.* Proc Natl Acad Sci U S A, 2007. **104**(11): p. 4730-5.

24. Davuluri, R., et al., *AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors.* BMC Bioinformatics, 2003. **4**(1): p. 25.

25. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Res, 2006. **34**(Database issue): p. D140-4.

26. Gustafson, A.M., et al., *ASRP: the Arabidopsis Small RNA Project Database.* Nucleic Acids Res, 2005. **33**(Database issue): p. D637-40.

27. Lu, C., et al., *Elucidation of the small RNA component of the transcriptome.* Science, 2005. **309**: p. 1525.

28. Masoudi-Nejad, A., et al., *EGENES: transcriptome-based plant database of genes with metabolic pathway information and expressed sequence tag indices in KEGG.* Plant Physiol, 2007. **144**(2): p. 857-66.

29. Zhang, P., et al., *MetaCyc and AraCyc. Metabolic pathway databases for plant research.* Plant Physiol, 2005. **138**(1): p. 27-37.

30. Gunsalus, K.C., et al., *Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis.* Nature, 2005. **436**(7052): p. 861-5.

31. Matthews, L.R., et al., *Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs".* Genome Res, 2001. **11**(12): p. 2120-6.

32. Gachon, C.M., et al., *Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis: functional and evolutionary implications.* Plant Mol Biol, 2005. **58**(2): p. 229-45.

33. Tatusov, R.L., E.V. Koonin, and D.J. Lipman, *A genomic perspective on protein families.* Science, 1997. **278**(5338): p. 631-7.

34. Zhang, J., *Evolution by gene duplication: an update.* TRENDS in Ecology and Evolution, 2003. **18**(6): p. 292-298.

35. Barrett, T. and R. Edgar, *Mining microarray data at NCBI's Gene Expression Omnibus (GEO)\*.* Methods Mol Biol, 2006. **338**: p. 175-90.

36. Pepper, S.D., et al., *The utility of MAS5 expression summary and detection call algorithms.* BMC Bioinformatics, 2007. **8**: p. 273.

37. Michaelewicz, Z. and D. Fogel, *How to Solve It: Modern Heuristics*. 2004: Springer Verlag.

38. Wang, L., et al., *A dynamic gene expression atlas covering the entire life cycle of rice.* Plant J, 2009.

39. Chiu, J.C., et al., *OrthologID: automation of genome-scale ortholog identification within a parsimony framework.* Bioinformatics, 2006. **22**(6): p. 699-707.

40. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic Acids Res, 2000. **28**(1): p. 33-6.

41. Berglund, A.C., et al., *InParanoid 6: eukaryotic ortholog clusters with inparalogs.* Nucleic Acids Res, 2008. **36**(Database issue): p. D263-6.

42. Murphy, K. and S. Mian, *Modelling Gene Expression Data using Dynamic Bayesian Networks*, in *Technical report, Computer Science Division*. 1999, University of California and Life Sciences Division, Lawrence Berkeley National Laboratory.

43. Wan, T.F., et al., *Correlation between AS1 gene expression and seed protein contents in different soybean (Glycine max [L.] Merr.) cultivars.* Plant Biol (Stuttg), 2006. **8**(2): p. 271-6.

44. Wang, R., et al., *Microarray analysis of the nitrate response in Arabidopsis roots and shoots reveals over 1,000 rapidly responding genes and new linkages to glucose, trehalose-6-phosphate, iron, and sulfate metabolism.* Plant Physiol, 2003. **132**(2): p. 556-67.

45. Gage, D.J., *Infection and invasion of roots by symbiotic, nitrogen-fixing rhizobia during nodulation of temperate legumes.* Microbiol Mol Biol Rev, 2004. **68**(2): p. 280-300.

46. Barbulova, A., et al., *Differential effects of combined N sources on early steps of the Nod factor-dependent transduction pathway in Lotus japonicus.* Mol Plant Microbe Interact, 2007. **20**(8): p. 994-1003.

47. Jeudy, C., et al., *Adaptation of Medicago truncatula to nitrogen limitation is modulated via local and systemic nodule developmental responses.* New Phytol, 2009.

48. Ruffel, S., et al., *Systemic signaling of the plant nitrogen status triggers specific transcriptome responses depending on the nitrogen source in Medicago truncatula.* Plant Physiol, 2008. **146**(4): p. 2020-35.

49. Gresshoff, P.M., *Post-genomic insights into plant nodulation symbioses.* Genome Biol, 2003. **4**(1): p. 201.

50. Bonneau, R., et al., *The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo.* Genome Biol, 2006. **7**(5): p. R36.

51. Barber, D., *Advances in Neural Information Processing Systems*. 2003, MIT Press, Cambridge MA. p. 729-736.

52. Rzhetsky, A., et al., *GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data.* Journal of Biomedical Informatics, 2004. **37**(1): p. 43-53.

53. Mueller, L.A., P. Zhang, and S.Y. Rhee, *AraCyc: A Biochemical Pathway Database for Arabidopsis.* Plant Physiol, 2003. **132**(2): p. 453.

54. Geisler-Lee, J., et al., *A predicted interactome for Arabidopsis.* Plant Physiol, 2007. **145**(2): p. 317-29.

55. Cui, J., et al., *AtPID: Arabidopsis thaliana protein interactome database--an integrative platform for plant systems biology.* Nucleic Acids Res, 2008. **36**(Database issue): p. D999-1008.

56. Mirowski P & LeCun Y (2009) Dynamical Factor Graphs for Time SeriesModeling. Lecture Note in Artificial Intelligence 5782:128-142

57. Gutierrez R, Gifford M, Poultney C, Wang R, Shasha D, Coruzzi G, Crawford N (2007) "Insights into the genomic nitrate response using genetics and the Sungear Software System" **Journal of Experimental Botany** doi: 10.1093/jxb/erm079

58. Efron B, Hastie T, Jonnstone I, Tibshirani R: **Least Angle Regression**. *Annals of Statistics* 2004, **32**(2):407-499.

59. Zhu H, Hastie T: **Regularization and variable selection via the elastic net**. *Journal of Royal Statistical Society* 2005, **67**:301-320.

60. Tibshirani R: **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistics Society* 2006, **58**(1):267-288.

61. Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, Thorsson V, Shannon P, Johnson MH, Bare JC *et al*: **A predictive model for transcriptional control of physiology in a free living cell**. *Cell* 2007, **131**(7):1354-1365.

62. Shimamura T, Imoto S, Yamaguchi R, Fujita A, Nagasaki M, Miyano S: **Recursive regularization for inferring gene networks from time-course gene expression profiles**. *BMC systems biology* 2009, **3**:41.

63. Wang Y, Joshi T, Zhang XS, Xu D, Chen L: **Inferring gene regulatory networks from multiple microarray datasets**. *Bioinformatics* 2006, **22**(19):2413-2420.

**APPENDIX-1 POSTDOCTORAL MENTORING PLAN**

**Co-mentorship across disciplines and institutions:** Post-Docs and students will receive novel cross-disciplinary training across biology & genomics (NYU Biology, Center for Genomics & Systems Biology) and informatics and systems Biology (NYU Courant). The PIs with expertise spanning these disciplines will co-mentor students and post docs individually and also at the weekly meetings where postdocs & students present their results.

**Training as Educators:** In this project, Post-Docs are paired up with graduate students, undergraduate students, and technicians in the laboratory/at the computer to practice mentoring skills in a research context. Post-docs are also afforded the opportunity to teach in NYU Biology courses where they are mentored by faculty advisors. For example, Dr. Katari, is currently co-teaching an undergraduate course “Introduction to Genomics & Bioinformatics” with a faculty mentor (Kris Gunsalus).

**Career Development:** Post-Docs receive counseling from their co-mentors and practice presentation skills during regular group-lab meetings, through a special NYU Biology Post-Doc seminar series, and at annual poster sessions at the NYU Biology retreat. Funds are provided for students and Post-Docs to attend at least one meeting each year and are expected to widely disseminate their work.

============= believe this goes elsewhere or nowhere ------

**Intellectual Property**

1. **Invention Disclosure and Patent Management:**

**Invention Disclosures** Invention disclosures will be reported by the inventors to the office responsible for patenting and licensing at NYU.

**Patenting** The institution at which the inventor is employed will be responsible for evaluating whether to retain title to inventions, filing U.S. and/or foreign patent applications, and notifying NSF. Inventions will be disclosed to the funding agency promptly upon receipt. Decisions on whether to file a patent application will be based upon an evaluation of the commercial potential of the invention by the institutional patenting and licensing office, and an evaluation of the patentability of the invention (including a search of prior art) in conjunction with an outside patent law firm. It is expected that the first patent filing will typically be a provisional filing. Decisions on whether to file non-provisional U.S. and or PCT applications will be made in the 10-11 month timeframe following the first (provisional) filing.

**Patent Reporting** NYU participates in the Interagency Edison System. NYU will notify the funding agency and grant the required non-exclusive license for government purposes for inventions for which NYU elects title. A final invention statement will be sent to NSF upon completion of the grant.

1. **Licensing and Commercialization:** The Office of Industrial Liaison at NYU shall be responsible for

seeking to make technology developed under the grant at NYU commercially available. Technology developed at the other participating institutions will be licensed by the technology licensing offices of those institutions. Research materials developed under the grant will be made available to researchers at not-for-profit institutions under a material transfer agreement with intellectual property terms consistent with NIH guidelines. The Office will also seek out appropriate industry partners to commercially develop technology. Potential arrangements will include exclusive or non-exclusive licenses. Any licenses will contain provisions requiring the licensing company to diligently develop the technology. Options may also be granted to companies considering licensing for limited time periods (e.g., 3-6 months) to allow them to conduct due diligence and evaluate their interest in a license. It is expected that technology will be licensed to existing companies, but if the technology is of sufficient breadth to justify the creation of a new start-up company, this will be considered.

=========== deleted text

**Gene Networks**: Gene Network analysis allows users to query our Gene Network data housed in the multinetwork and displays the results in a graph using Cytoscape, an open source project for which we have built a plug-in. The tool allows users to filter interactions before displaying the graph {Gutierrez, 2007 #12}.

**Supernode Networks**: The Supernode Network helps summarize the results of a Gene Network analysis. Individual genes in the gene network are grouped (Supernode) on the basis of their functional annotations and they are associated with other Supernodes with edges determined from the Gene Network data. A Supernode’s size is determined by the number of genes that it contains.

**Creating a state-space regulatory network for Rice.** We will use a similar time series-based approach to examine nitrogen regulatory networks in Rice. For this, we will use similar growth and N-treatments in Rice that we have used in Arabidopsis, and for which we have shown Nitrogen-regulation of gene expression in Rice. Rice seedlings (approximately 14 days old) will be grown in nitrate-free sterile hydroponic media containing 0.5 mM NH4+ succinate. Plants will subsequently be N-deprived for 24 hours, then treated with 1 mM KNO3- and then shoot and root tissue will be harvested at 0, 3,6, 9, 12, 15, 20, 25, 35, 45, 60, 120, 180 and 240 min. Controls are "T0" (harvest time zero, before treatment) and 1 mM KCl at each of these time-points. To select time-points for transcriptome analysis, "sentinel" NO3- regulated genes involved in early steps of the nitrate response (e.g. nitrate uptake or reduction) will be monitored by Q-PCR {Wan, 2006 #43;Wang, 2003 #44;Wang, 2004 #11}. Based on these Q-PCR results, an initial set of RNA time points will be analyzed using transcriptomics in biological duplicates. We will prepare mRNA samples for RNA-seq (Illumina) using protocols we have previously used for Arabidopsis. Briefly, polyA mRNA will be sheared and sized to ~50-60 bp and ligated to adaptors at 5’ and 3’ ends for reverse transcription and PCR amplification. 14 cycles will be used for enrichment, to minimize PCR bias, but produce enough material for Illumina sequencing (see McCombie letter, CHSL).

**Step 1. Identify the “regulated" genes in the target species (Rice).** To find the genes that are N-regulated in Rice in our time-series data, we will use criteria similar to those used for Arabidopsis to minimize the False Discovery Rate (FDR). As we did for Arabidopsis time-series data {Krouk, 2010 #19}, we will run an ANOVA (aov() function) over the data set where the signal of a probe i is Pi ~ µ + αN +βT+ γT\*N + ε where N is the effect of the Nitrate treatment, T is the effect of time, and T\*N is the effect of their interaction, µ is the mean signal over the data set, and ε the unexplained variance. We determine for each gene the particular time point that showed the most marked effect of nitrate. We call a probe “regulated” if it has a positive call (FDR <5%) of interaction with nitrate at that particular time point. This will lead us to the identification of a nitrogen-regulated network having regulators such as transcription factors (TFs) and targets of regulation (all genes – both targets and TFs).

**Step 2. Find inferred regulatory and homology-supported correlation edges between Arabidopsis and Rice.** As a starting point for further analysis, we will next find edges between Rice genes (e.g. R1.1, R2.2) having strong Pearson correlation, where Rice gene R1.1 is a transcription factor, for which binding site over-representation also gives support {Gutierrez, 2008 #7;Nero, 2009 #8}, and such that homologous genes in Arabidopsis (A1, A2) also comprise a putative edge in the transcription factor network generated from the time series data. We seek Pearson correlation between Rice genes having p-value < 0.01 (that is a 0.01 chance that the correlation is due to chance). We will use these inferred TFTargetTarget regulatory edges among Rice genes to “prime the pump” in the “State-Space Modeling” analysis of the next step.

**Step 3. Infer the regulatory network using “State Space” analysis.** Using the homology-aided correlation edges from Step 2, we set the initial weights of the TFargettarget pairs for the State-Space modeling approach {Mirowski, 2009 #56}. Specifically, gene pairs that are positively or negatively correlated in the analysis of Step 2, will have large (positive or negative, respectively) weights. As State-Space modeling is an iterative algorithm, those weights will be adjusted in the course of error reduction. Better initial edge weights can improve performance. For example, in our Arabidopsis study, the knowledge that certain genes were transcription factors improved the prediction accuracy of State Space analysis by 5% (from 67% to nearly 71%) (Fig. 5B) {Krouk, 2010 #19}.

**Step 4. Validate the *inferred* Rice regulatory network and plan new experiments.** Validation here will mean prediction of Rice regulatory networks under untested conditions, and validation of these predictions using experimental data. Just as we did for Arabidopsis, we will build a network with the Rice time series transcriptome data, but leaving out certain time points, and then predict the values at those missing time points and compare the predicted values with the actual ones. If the network predictions are not statistically significantly better than random based on p-value analysis, we will analyze more time-series experiments. To decide which new time series data points to use, we will make use of the following simple heuristic: perform a new measurement of a sample that is most similar to the most useful measurement just done as judged by prediction accuracy. For example, we can determine which of the existing samples are most valuable by removing them and computing prediction accuracy. In the case of our Arabidopsis time-course study {Krouk, 2010 #19}, removing two replicate experiments from two different time-points prior to 15 minutes is less harmful to the accuracy of the prediction of the network state at 20 min than removing both replicates from a single time-point prior to 15 min. This suggests that measurements at different time-points are more valuable than replicates.

11

**Operation of Cross-Species Network Inference (CSNI) in *Inference* mode:**

**Step 1.** **Choose a target (crop), reference species and Homology algorithm**. On the web site (see a mockup of the CSNI GUI in Fig. 6), the plant scientist selects the reference species and the target species. Any species can conceivably serve as the reference species provided: i) it is phylogenetically close to the target species (even as distant as Rice-Arabidopsis); and ii) it has data to support the construction of a validated network. The plant scientist next chooses a homology algorithm and its parameters, such as BLAST and its E-value cutoff, or a parsimony mechanism, and CSNI generates a set of homologous reference-to-target gene pairs. Preliminary tests in Aim 1, showed that reverse top hits gave high precision for Arabidopsis (as reference) and Rice (as target). Further work will determine whether this is the best distanced-based homology strategy and how this compares to parsimony based homology methods. In addition, we will allow users to upload any cross-species homologous gene pairs determined using their preferred method (e.g. COG {Tatusov, 2000 #40}, InParanoid {Berglund, 2008 #41} OrthologID {Chiu, 2006 #39}, or homology pairs generated using analysis platforms such as Taverna {Oinn, 2004 #2}, Kepler {Altintas, 2004 #3}, or Galaxy {Blankenberg, #4}, for example).

In our working example, we use Arabidopsis as our reference species for Medicago, because the Arabidopsis genome contains a large number of nodulin-like genes {Gresshoff, 2003 #49}. Moreover, our analysis showed that of a list of 1,458 potential Arabidopsis nodulation gene homologs (collated from Allometra database (<http://allometra.com/nodprots.shtml>)), 36% of these genes were N-regulated; the majority were N-depressed, consistent with externally supplied N inhibiting nodulation in legumes. As early nodulation events are inhibited by external nitrogen application, potentially homologous genes in Arabidopsis and Medicago and their associated N-regulated gene networks could provide insights into N-regulated root development and to the events involved in N-repression of nodulation and N-fixation in Medicago.

**Step 2**. **Obtain and analyze experimental data in the target species.** In this step, the plant scientist gathers and/or conducts experiments about the target plant (e.g. transcriptome). For purposes of analysis, the pipeline will offer standard tools such as correlation, linear regression, as well as machine learning tools such as “State Space” analysis (Aim 2) to identify such relationships.

**Step 3**. **Infer a network in the target species.** The plant scientist next uses the data obtained in Steps 1 and 2, plus the CSNI inference engine and its parameters to infer a putative network in the target species, as depicted in Fig. 3A, as follows. First, the biologist chooses a validated network in the reference species. The biologist also selects which types of edges should be included in the reference species’ validated network (e.g. protein:protein, metabolic, etc), as well as some parameter settings or Inference rules that are generated using an optimization technique such as simulated annealing as discussed in Aim 1, Step 7. For example, a combination rule might infer a regulatory edge in the target species if the edge’s genes were connected by an expression edge with correlation > 0.7 and the edge had homologous genes connected by a regulatory edge in the reference species’ validated network. CSNI next takes 1) the homology mapping between the reference and target species chosen in Step 1, and 2) the validated network in the reference species, and 3) the set of experimental data about the target species chosen. In our case, we would use the time series data for Medicago from Aim 3, as well as already existing transcriptome data in Medicago (<http://bioinfo.noble.org/gene-atlas/v2/>). Given this data, CSNI infers a gene network for the target species.

We will implement CSNI as a general-purpose tool to be used by plant scientists. We plan to deploy CSNI ([www.CrossSpecies.org](http://www.CcrossSspecies.org)) on several additional platforms, first on our VirtualPlant website (www.virtualplant.org), and second on *iPlant* (see S. Goff letter). We will also deploy CSNI on one of the widely-used bioinformatic workflow engines: Taverna {Oinn, 2004 #2}, Kepler {Altintas, 2004 #3} or Galaxy {Blankenberg, #4}. Implementing the CSNI pipeline on top of one (or more) of these bioinformatic workflow engines is important because they provide increasingly popular platforms for developing computational genetic analyses, and provide generic support for reproducible bioinformatic analyses.

While these three steps comprise the basic production operation of CSNI in Inference mode, we will also investigate three related enhancements.

**Enhancements to the Operation of CSNI in *Inference* mode:**

**A.** **Select multiple reference species.** Some target species may be phylogenetically close to multiple possible reference species. In this situation, the plant scientist may execute CSNI repeatedly with different reference species and then combine the resulting inferred networks. The networks may be combined based on set union or intersection functions or based on more sophisticated weighting functions determined by biological intuition or learning as in Aim 1, Step 7. These functions will be enhancements to the CSNI pipeline.

**B.** **Assemble validated networks for potential target species.** We will assemble validated networks for Grape and Corn (for which KEGG pathways exist) and in Medicago we will infer metabolic pathways, until a KEGG version is available. These validated networks will help users validate inferred networks, and help set CSNI parameters through optimization studies in these crop species based on similarity scores and p-values computed using ***NetSect***, as described in Aim 1. Time and resources permitting, we will also enable validated networks for additional crop species, in consultation with **iPlant.**

**C. Suggest the next experiment.** As discussed in Aim 2, Step 4, in State-Space modeling a simple but powerful method for determining the most useful previous experiment or set of experiments is to suppose those experiments didn’t exist and then measure how much that degrades accuracy predictions. Using that information can help a biologist determine the next measurement to use (e.g. a new time point in a time series rather than an additional replicate). This is meant as an aid rather than as a substitute for biological insight.

**Expected outcomes of Aim 4.** The CSNI pipeline analysis constructed in Aim 4, and made available to the community as a biologist-friendly interface, will empower plant biologists to use network approaches to derive testable hypothesis for gene functions in crop species for which limited genomic information is available. Identifying networks conserved between reference and crop species will also enable researchers to focus their translational studies from models to crops.